UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

| APPLICATION NO. | FILING DATE | FIRST NAMED INVENTOR | ATTORNEY DOCKET NO. | CONFIRMATION NO. |
|---|---|---|---|---|
| 10/662,985 | 09/15/2003 | Min Chu | M61.12-0565 | 2246 |

27366 7590 12/19/2007

WESTMAN CHAMPLIN (MICROSOFT CORPORATION)
SUITE 1400
900 SECOND AVENUE SOUTH
MINNEAPOLIS, MN 55402-3319

| EXAMINER | |
|---|---|
| COLUCCI, MICHAEL C | |

| ART UNIT | PAPER NUMBER |
|---|---|
| 2626 | |

| MAIL DATE | DELIVERY MODE |
|---|---|
| 12/19/2007 | PAPER |

**Please find below and/or attached an Office communication concerning this application or proceeding.**

The time period for reply, if any, is set in the attached communication.

| Office Action Summary | Application No. | Applicant(s) | |
|---|---|---|---|
| | 10/662,985 | CHU ET AL. | |
| | Examiner | Art Unit | |
| | Michael C. Colucci | 2626 | |

*-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --*

**Period for Reply**

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE *3* MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

**Status**

1)☐ Responsive to communication(s) filed on _____.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

**Disposition of Claims**

4)☒ Claim(s) *1-30* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1-30* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

**Application Papers**

9)☐ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on *15 September 2003* is/are: a)☒ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

**Priority under 35 U.S.C. § 119**

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

**Attachment(s)**

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☒ Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date *12/[illegible]/04, 08/22/05, 12/29/06*.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____ .
5) ☐ Notice of Informal Patent Application
6) ☐ Other: _____.

## DETAILED ACTION

### *Claim Rejections - 35 USC § 102*

1. The following is a quotation of the appropriate paragraphs of 35 U.S.C. 102 that form the basis for the rejections under this section made in this Office action:

> A person shall be entitled to a patent unless –
>
> (b) the invention was patented or described in a printed publication in this or a foreign country or in public use or on sale in this country, more than one year prior to the date of application for patent in the United States.

2. Claims 1-11, 16-18, 20-21, 23, 26-28, and 30 rejected under 35 U.S.C. 102(b) as being anticipated by Huang et al "Recent improvements on Microsoft's trainable text-to-speech system-Whistler" (hereinafter Huang).

Re claim 1, Huang teaches a method for synthesizing speech, the method comprising:

generating a training context vector for each of a set of training speech units (Abstract) in a training speech corpus, each training context vector indicating the prosodic context of a training speech unit in the training speech corpus (page 960 col 1 paragraph 3);

indexing a set of associated with a speech units based speech segments set of training (page 961 col 2 paragraph 2) on the context vectors for the training speech units (page 960 col 2 *Prosody contour generation*);

generating an input context vector for each of a set of input speech units in an input text (page 960 col 2 *Clause specification*), each input context vector indicating the

prosodic context of an input speech unit in the input text (page 960 col 2 paragraph 3-4);

using the input context vectors to find a speech segment for each input speech unit (page 961 col 1 3.1 Unit Generation);

concatenating the found speech segments to form a synthesized speech signal (page 961 col 1 3.1 Unit Generation).

Re claim 2, Huang teaches the method of claim 1 wherein the each context vector comprises a position-in-phrase coordinate (page 960 col 1 2.2 Prosody Model) indicating the position of the speech unit in a phrase (page 962 col 1 4.0 Summary).

Re claim 3, Huang teaches the method of claim 1 wherein the each context vector comprises a position-in-word coordinate (page 960 col 1 2.2 Prosody Model) indicating the position of the speech unit in a word (page 962 col 1 4.0 Summary).

Re claim 4, Huang teaches the method of claim 1 wherein the each context vector comprises a left phonetic coordinate indicating a category for the phoneme to the left of the speech unit (page 961 col 2 paragraph 1).

Re claim 5, Huang teaches the method of claim 1 wherein the each context vector comprises a right phonetic coordinate indicating a category for the phoneme to the right of the speech unit (page 961 col 2 paragraph 1).

Re claim 6, Huang teaches the method of claim 1 wherein the each context vector comprises a left tonal coordinate indicating a category for the tone (page 961 col 1 paragraph 3) of the speech unit to the left of the speech unit (page 961 col 2 paragraph 1).

Re claim 7, Huang teaches the method of claim 1 wherein the each context vector comprises a right tonal coordinate indicating a category for the tone (page 961 col 1 paragraph 3) of the speech unit to the right of the speech unit (page 961 col 2 paragraph 1).

Re claim 8, Huang teaches the method of claim 1 wherein the each context vector comprises a coordinate indicating a coupling degree of pitch, duration and/or energy with a neighboring unit (page 960 col 1 2.1 Text Analysis Model).

Re claim 9, Huang teaches the method of claim 1 the each context vector comprises a coordinate indicating a level of stress of a speech unit (page 960 col 2 *Clause Specification*).

Re claim 10, Huang teaches the method of claim 1 wherein indexing a set of speech segments comprises generating a decision tree based on the training context vectors (page 961 col 2 paragraph 1).

Re claim 11, Huang teaches the method of claim i0 wherein using the input context vectors to find a speech segment comprises searching the decision tree using the input context vector (page 961 col 2 paragraph 1).

Re claim 16, Huang teaches the method of claim 1 wherein the context vector comprises one or more higher order coordinates being combinations of at least two factors from a set of factors including:

an indication of a position of a speech unit in a phrase;

an indication of a position of a speech unit in a word;

an indication of a category for a phoneme preceding a speech unit (page 961 col 2 paragraph 1);

an indication of a category for a phoneme following a speech unit (page 961 col 2 paragraph 1);

an indication of a category for tonal identity of the current speech unit;

an indication of a category for tonal identity of a preceding speech unit;

an indication of a category for tonal identity of a following speech unit;

an indication of a level of stress of a speech unit;

an indication of a coupling degree of pitch, duration and/or energy with a neighboring unit; and an indication of a degree of spectral mismatch with a neighboring speech unit.

Re claim 17, Huang teaches a method of selecting sentences for reading into a training speech corpus synthesis used in speech (page 960 col 1 paragraph 3), the method comprising:

identifying a set of prosodic context information for each of a set of speech units (page 960 col 1 paragraph 3);

determining a frequency of occurrence for each distinct context vector that appears in a very large text corpus (page 961 col 1 paragraph 2-3);

using the frequency of occurrence of the context vectors to identify a list of necessary context vectors (page 961 col 1 paragraph 2-3); and

selecting sentences in the large text corpus for reading into the training speech corpus, each selected sentence containing at least one necessary context vector (page 961 col 1 paragraph 2-3).

NOTE: A tonal pattern relevant to speech is construed to be both functionally equivalent and effective to prosodic context information, where intonation is a part of prosodic information.

Re claim 18, Huang teaches the method of claim 17 wherein identifying a collection of prosodic context information sets as necessary context information sets comprises:

determining the frequency of occurrence of each prosodic context information set across a very large text corpus (page 961 col 1 paragraph 2-3);

identifying a collection of prosodic context information sets as necessary context information sets based on their frequency of occurrence (page 961 col 1 paragraph 2-3).

Re claim 20, Huang teaches the method of claim 17 further comprising indexing only those speech segments that are associated with sentences in the smaller training text (page 961 col 1 paragraphs 2-3) and wherein indexing comprises indexing using a decision tree (page 961 col 2 paragraph 1-2).

Re claim 21, Huang teaches the method of claim 20 wherein indexing further comprises indexing the speech segments in the decision tree based on information in the context information sets (page 961 col 2 paragraph 1-2).

Re claim 23, Huang teaches a method of selecting speech segments for concatenative speech synthesis (page 961 col 1 3.1 Unit Generation) the method comprising:

parsing an input text into speech units (page 961 col 2 paragraph 1-2);

identifying context information for each speech unit based on its location in the input text and at least one neighboring speech unit (page 961 col 2 paragraph 1);

identifying a set of candidate speech segments for each speech unit based on the context information (page 960 col 2 *Prosody contour generation*);

identifying a sequence of speech segments from the candidate speech segments (page 960 col 2 *Prosody contour generation*) based in part on a smoothness cost between the speech segments (page 960 col 2 *Stochastic variation* & *Contour Interpolation and Smoothing*).

Re claim 26, Huang teaches the method of claim 23 wherein identifying a sequence of speech segments comprises using a smoothness cost (page 960 col 2 *Stochastic variation* & *Contour Interpolation and Smoothing*) that is based on whether two neighboring candidate speech segments appeared next to each other in a training corpus (page 961 col 2 paragraph 1).

NOTE: For purposes of prior art a smoothness cost is construed to be functionally equivalent and effective as a contour interpolation an smoothing component, used to create a more natural effect. Huang teaches a natural speech synthesis, where if an exact match is not found, a cost function is employed to synthesis the most natural speech.

Re claim 27, Huang teaches the method of claim 23 wherein identifying a sequence of speech segments comprises using an objective measure comprising one or more first order components from a set of factors comprising:

an indication of a position of a speech unit in a phrase;

an indication of a position of a speech unit in a word;

an indication of a category for a phoneme preceding a speech unit (page 961 col 2 paragraph 1);

an indication of a category for a phoneme following a speech unit (page 961 col 2 paragraph 1);

an indication of a category for tonal identity of the current speech unit;

an indication of a category for tonal identity of a preceding speech unit;

an indication of a category for tonal identity of a following speech unit;

an indication of a level of stress of a speech unit;

an indication of a coupling degree of pitch, duration and/or energy with a neighboring unit;

an indication of a degree of spectral mismatch with a neighboring speech unit.

Re claim 28, Huang teaches the method of claim 23 wherein identifying a sequence of speech segments comprises using an objective measure comprising one or more higher order components being combinations of at least two factors from a set of factors including:

an indication of a position of a speech unit in a phrase;

an indication of a position of a speech unit in a word;

an indication of a category for a phoneme preceding a speech unit (page 961 col 2 paragraph 1);

an indication of a category for a phoneme following a speech unit (page 961 col 2 paragraph 1);

an indication of a category for tonal identity of the current speech unit;

an indication of a category for tonal identity of a preceding speech unit;

an indication of a category for tonal identity of a following speech unit;

an indication of a level of stress of a speech unit;

an indication of a coupling degree of pitch, duration and/or energy with a neighboring unit;

an indication of a degree of spectral mismatch with a neighboring speech unit.

Re claim 30, Huang teaches a computer-readable medium having computer executable instructions for synthesizing speech from speech segments based on speech units found in an input text, the speech being synthesized through a method comprising steps of:

identifying context information for each speech unit based on the prosodic structure of the input text (page 961 col 2 paragraph 1-2);

identifying a set of candidate speech segments for each speech unit based on the context information (page 960 col 2 *Prosody contour generation*);

identifying a sequence of speech segments from the candidate speech segments (page 960 col 2 *Prosody contour generation*);

concatenating the sequence of speech segments without modifying the prosody of the speech segments to form the synthesized speech (page 959 col 1 – col 2 Introduction & Fig. 1 Synthesis Phase).

***Claim Rejections - 35 USC § 103***

3. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

4. Claims 12-15, 22, 24-25, and 29 rejected under 35 U.S.C. 103(a) as being unpatentable over Huang et al "Recent improvements on Microsoft's trainable text-to-speech system-Whistler" (hereinafter Huang) in view of Seide US 5857169 A (hereinafter Seide).

Re claim 12, Huang teaches method of claim 11 wherein searching the decision tree (page 961 col 2 paragraph 1) comprises:

The selection is based on a cost function (page 961 col 1 paragraph 1-2).

However Huang fails to teach identifying a leaf in the tree for each input context vector, each leaf comprising at least one candidate speech segments (Seide col 3 line 9-35 & Fig. 4);

selecting one candidate speech segment in each leaf node, wherein if there is more than one candidate speech segment on the node (Seide col 3 line 9-35 & Fig. 4).

Seide teaches organizing the reference probability densities, using a tree structure, and performing a tree search. At the lowest level of the tree (level 1), each of the leaf nodes corresponds to an actual reference probability density. Seide teaches that a reference probability density represents an elementary cluster of reference

vectors. At level two of the tree, each non-leaf node corresponds to a cluster probability density.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention identifying a leaf in a tree for an input context vector, where a candidate is selected and if multiple candidates exist, a cost function is implemented. Using a cost function when searching a tree would allow for a minimized cost and maximized probability of a match, where a mismatch would be less likely to occur. Since the decision tree would have to be searched regardless, if another potential candidate occurs, by incorporating a cost function for similarity comparison the distance between vectors would be decreased.

Re claim 13, Huang fails to teach the method of claim 12 wherein the cost function comprises a distance between the input context vector and a training context vector associated with a speech segment (Seide col 10 line 6-13).

Seide teaches a key step in locating a reference pattern which corresponds to the input pattern is finding the reference vector which is `nearest` the observation vector (nearest neighbor search), where `distance` is defined as the negative logarithm of the likelihood.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention finding a distance between a training context vector and a speech segment. Using a distance between vectors would allow for a cost calculation having a

directly proportional value of cost relevant to distance between vectors, where a minimal distance would be desirable, yielding the highest probability.

Re claim 14, Huang teaches the method of claim 13 wherein the cost function (page 961 col 1 paragraph 1-2) further comprises a smoothness cost (page 960 col 2 *Stochastic variation & Contour Interpolation and Smoothing*) that is based on a candidate speech segment of at least one neighboring speech unit (page 961 col 2 paragraph 1).

NOTE: For purposes of prior art a smoothness cost is construed to be functionally equivalent and effective as a contour interpolation an smoothing component, used to create a more natural effect. Huang teaches a natural speech synthesis, where if an exact match is not found, a cost function is employed to synthesis the most natural speech.

Re claim 15, Huang teaches the method of claim 14 wherein the smoothness cost (page 960 col 2 *Stochastic variation & Contour Interpolation and Smoothing*) gives preference to selecting a series of speech segments for a series of input context vectors if the series of speech segments occurred in series in the training speech corpus (page 961 col 1 3. Unit Generation & col 2 paragraph 1-2).

Re claim 22, Huang fails to teach the method of claim 21 wherein the decision tree comprises leaf nodes and at least one leaf node comprises at least two speech segments for the same speech unit (Seide col 3 line 9-35 & Fig. 4).

Seide teaches organizing the reference probability densities, using a tree structure, and performing a tree search. At the lowest level of the tree (level 1), each of the leaf nodes corresponds to an actual reference probability density. Seide teaches that a reference probability density represents an elementary cluster of reference vectors. At level two of the tree, each non-leaf node corresponds to a cluster probability density. Seide also teaches training observation vectors characterized by two cluster in Fig. 4, where each node will break into two clusters of data.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention a decision tree having multiple nodes with a leaf of two or more speech segments. It is obvious to use a tree to segment text. Using a decision tree would produce a specific segmentation of text such as diphones and triphones, rather than just segmentation of a sentence into words, a training system can further classify speech segments at a sub-phonetic level to have a more flexible memory quality.

Re claim 24, Huang teaches the method of claim 23 wherein identifying a set of candidate speech segments for a speech unit comprises applying the context information for a speech unit to a decision tree (page 961 col 2 paragraph 1).

However Huang fails to particularly teach to identify a leaf node containing candidate speech segments for the speech unit (Seide col 3 line 9-35 & Fig. 4).

Seide teaches organizing the reference probability densities, using a tree structure, and performing a tree search. At the lowest level of the tree (level 1), each of the leaf nodes corresponds to an actual reference probability density. Seide teaches that a reference probability density represents an elementary cluster of reference vectors. At level two of the tree, each non-leaf node corresponds to a cluster probability density.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention identifying a leaf node in a tree for an input context vector, where a candidate is selected and if multiple candidates exist, a cost function is implemented. Using a cost function as part of a decision tree to determine differences while searching a tree, would allow for a minimized cost and maximized probability of a match, where a mismatch would be less likely to occur. Since the decision tree would have to be searched regardless, if another potential candidate occurs, by incorporating a cost function for similarity comparison the distance between vectors would be decreased.

Re claims 25 and 29, Huang fails to teach the method of claim 24 wherein identifying a set of candidate speech segments further comprises pruning some speech segments from a leaf node based on differences between the context information of the speech unit from the input text and context information associated with the speech segment (Seide col 4 line 20-39 & Fig. 4).

Seide teaches organizing the reference probability densities, using a tree structure, and performing a tree search. At the lowest level of the tree (level 1), each of

the leaf nodes corresponds to an actual reference probability density. Seide teaches that a reference probability density represents an elementary cluster of reference vectors. At level two of the tree, each non-leaf node corresponds to a cluster probability density. Seide also teaches the subset with the highest likelihood, i.e., the `nearest` to the observation vector, is determined and if the difference of the highest likelihood and the likelihood of the other subset is below a threshold, then also the other subset is computed further. In this way, the number of reference probability densities, which are finally selected at level one of the tree, can be dynamically determined. Seide demonstrates the pruning of portions of a tree in Figure 4, where a segments are simplified further to more specific elements.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention pruning speech segments from a leaf node based on differences between context information and input text. Pruning a tree with respect to differences would allow for the highest likelihood to be the resultant node on a tree, which would minimize a cost function and produce the most natural sounding speech.

**5. Claim 19 rejected under 35 U.S.C. 103(a) as being unpatentable over Huang et al "Recent improvements on Microsoft's trainable text-to-speech system-Whistler" (hereinafter Huang) in view of Ahmad et al US 6172675 B1 (hereinafter Ahmad).**

Re claim 19, Huang fails to teach method of claim 18 wherein identifying a collection of prosodic context information sets as necessary context information sets further comprises:

sorting the context information sets by their frequency of occurrence in decreasing order (col 10 line 47 – col 11 line 9);

determining a threshold, F, for accumulative frequency of top context vectors (col 10 line 47 – col 11 line 9);

selecting the top context vectors whose accumulative frequency is not smaller than F for each speech unit as necessary prosodic context information sets (col 10 line 47 – col 11 line 9).

Ahmad teaches that words longer than a specified threshold length are termed "complex." Complex words that appear in the text data set, except those listed in a "stop list" of words deemed to be insufficiently specific to the topic at hand (an initial stop list includes general, common words in the text's language) are identified and sorted according to frequency of occurrence. The most frequently occurring complex words are identified as "seed" words. Ahmad teaches that the previously identified seed words can be added to the stop list (since they apparently were not sufficiently selective) and the threshold length used to identify complex words can be decreased so that more seed words can be identified.

Therefore, it would have been obvious to one of ordinary skill in the art at the time of the invention sorting context information in decreasing order by frequency of occurrence, where a threshold is determined to select the top context vectors not

smaller than the threshold. Using a threshold for the sorting of context information based on the frequency of occurrence would allow for identifying and distinguishing between common words and words that relate to a text summarization (i.e. relevant/matching data). This would create a more specific and efficient way of classifying data as a system is trained.

***Conclusion***

6. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure US 4718094 A, US 5912989 A, US 5933806 A, US 5146405 A, US 5937422 A, US 5440481 A.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Michael C. Colucci whose telephone number is (571)-270-1847. The examiner can normally be reached on 9:30 am - 6:00 pm, Monday-Friday.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Richemond Dorvil can be reached on (571)-272-7602. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

RICHEMOND DORVIL
SUPERVISORY PATENT EXAMINER

Michael Colucci Jr.
Patent Examiner
AU 2626
(571)-270-1847
Michael.Colucci@uspto.gov